Science Journal of Applied Mathematics and Statistics
2019; 7(3): 45-50
http://www.sciencepublishinggroup.com/j/sjams
doi: 10.11648/j.sjams.20190703.13
ISSN: 2376-9491 (Print); ISSN: 2376-9513 (Online)

Application of Association Rule Mining in Talent Introduction Analysis

Zixuan Chen, Jiepin Ding*, Zhiguang Zhou, Yin Zhu, Wenyu Zhang

School of Information, Zhejiang University of Finance and Economics, Hangzhou, China

Email address:

chenzx@zufe.edu.cn (Zixuan Chen), dingjp@zufe.edu.cn (Jiepin Ding), zhgzhou1983@163.com (Zhiguang Zhou), valentine-zy@zufe.edu.cn (Yin Zhu), wyzhang@e.ntu.edu.sg (Wenyu Zhang)

*Corresponding author

To cite this article:

Zixuan Chen, Jiepin Ding, Zhiguang Zhou, Yin Zhu, Wenyu Zhang. Application of Association Rule Mining in Talent Introduction Analysis. *Science Journal of Applied Mathematics and Statistics.* Vol. 7, No. 3, 2019, pp. 45-50. doi: 10.11648/j.sjams.20190703.13

Received: August 5, 2019; **Accepted**: September 25, 2019; **Published**: September 27, 2019

Abstract: With the advancement of higher education, many colleges have given increasing attention to talent introduction. On the other hand, the association rule mining technique is a useful method which extracts the useful association rules from the complex data repositories. This study takes the example of 245 academic staff from Zhejiang University of Finance and Economics, China and uses Apriori algorithm to explore the association rules on whether an academic staff can obtain the Natural Science Foundation of China (NSFC) within three years after s/he is recruited to the university. The aim of this study is to better introduce talents for colleges so that the academic levels of colleges can be improved. The results of association rule mining have shown that having published high quality papers such as SCI paper and SSCI paper has an important effect on the probability of academic staff to obtain NSFC within three years. Besides, the grade of PhD school has also an effect on the probability of academic staff to obtain NSFC within three years. The higher the grade of a staff's PhD school is, the easier for him to obtain NSFC within three years.

Keywords: Data Mining, Association Rule Mining, Apriori Algorithm

1. Introduction

The talent is crucial for the development of economy and society, and more and more organizations have paid attention to talent recruitment and selection. Therefore, many studies have been conducted on talent selection to improve the decision making effect in human resource management. Especially, to increase the level of higher education, many colleges have given increasing attention to talent introduction. However, the existing evaluation approaches of talent selection are defined via specific post and duties according to their static characteristics [1]. Furthermore, there are very few studies on effective approaches conducted on talent introduction for colleges. Hence, it is believed that the level of higher education can be improved if the problem of talent introduction for colleges can be solved by effective approach.

Data mining is a data processing technique for extracting hidden knowledge from a series of complex data, by means of association rule mining method, prediction method, classification method, and clustering method [2, 3]. In particular, the development of association rule mining provides a method to extract causal structures and interesting relationships from the sets of databases. In order to better introduce talents for colleges, this study takes the example of 245 academic staff from Zhejiang University of Finance and Economics, China and uses the index indicating whether an academic staff can obtain the Natural Science Foundation of China (NSFC) within three years after s/he is recruited to the college to evaluate the staff's academic capacity. The method of association rule mining and Apriori algorithm are employed to find the correlation between whether an academic staff can obtain NSFC within three years and personal information. The R programming is used to do the experiment of association rule mining.

The rest of this study is organized as follows. Section 2 introduces the related works of association rule mining and Apriori algorithm. Section 3 gives a detailed descriptions of data preprocessing and association rule mining. Section 4 mainly analyses the results of association rule mining. The

conclusion of this study and our future work are introduced in the last section.

2. Related Work

With the arrival of big data era, the technique of data mining is becoming increasingly important. As one of the important data mining methods, association rule mining has been studied by many researchers. Lin et al. [4] employed adaptive-support association rule mining to find associations between users as well as associations between items for recommender systems. Qodmanan et al. [5] proposed a flexible association rule mining method based on genetic algorithm, which used multi-objective fitness instead of support and confidence value to evaluate rules. Huang et al. [6] proposed a new cloud-assisted association rule mining algorithm, which can minimize risks of privacy leakage. Beiranvand et al. [7] employed multi-objective particle swarm optimization algorithm and multi-objective perspective to solve the numerical association rule mining problem. Gyenesei [8] combined fuzzy set theory with association rule mining to find fuzzy association rules. Fung et al. [9] combined the association rule mining method with multi-objective genetic algorithm in order to discover a set of customer evaluation rules that can calculate the lower and higher limits of the design patterns. Alatas et al. [10] proposed a search strategy of multi-objective differential evolution approach for exploring accurate and optimal association rules, subject to the condition that all objectives are simultaneously included.

The development of Apriori algorithm [11] provides a method for association rule mining. Guo et al. [12] used improved Apriori algorithm for the problem of mobile e-commerce shopping, which is aimed to avoid information overload. Li et al. [13] employed Apriori algorithm for association rule mining to explore the relationships between stroke risk factors. Guo et al. [14] proposed a new model of wind speed forecasting based on time series, in which, Apriori algorithm is used to find the association rules. Singh et al. [15] proposed an extended Apriori algorithm in order to reduce the scanning time, by means of removing unnecessary transaction records and redundant generation of sub-items. Yu et al. [16] employed improved Apriori algorithm to solve the bottleneck problems based on the Boolean matrix.

However, only few studies have been conducted on exploring the relationship between talent capacity and personal information. Furthermore, the researches on exploring those relations mentioned above based on data mining are also rare. This study aims to explore the relationship between staff's academic capacity and personal information. Whether an academic staff can obtain NSFC within three years after s/he is recruited to the university is used as an indicator to evaluate the staff's academic capacity. The results of this study provide a reference for colleges to recruit the academic talents.

3. Data Preprocessing and Association Rule Mining

This section elaborates on the detailed descriptions of data preprocessing and association rule mining.

3.1. Removal of Invalid Attributes and Records

Because the raw data are incomplete and some values of attributes for most staff are similar, which would affect the accuracy of the association rule mining results, the invalid attributes and records are removed before doing association rule mining.

Firstly, because the aim of this study is exploring the relationship between whether an academic staff can obtain NSFC within three years and personal attributes, the records whose staff employment period is less than three years should be removed. Secondly, considering that most staff are full-time teachers and have obtained doctorate degree, the attributes of the highest degree of education, appointment post, and full-/part-time property are removed. Thirdly, the larger the number of attributes is, the more complex the problem solved via association rule mining is, which would affect the accuracy of the association rule mining results. Therefore, some less important attributes such as undergraduate time, and graduation time of doctorate after running experimental tests are removed. Finally, 172 records that contain 17 personal attributes of staff are obtained, such as age, no. of published SCI paper, and no. of published SSCI paper.

3.2. Addition of Derived Attributes

In this study, in order to explore the relationship between no. of published paper and whether NSFC can be obtained within three years, an additional indicator Scientific Research Ability (SRA) is proposed, which represents the overall number and grade of papers published by each academic staff. Because the SRA is not only related to the number of papers but also the grade of papers, this study gives different grade of papers a certain weight to calculate the SRA. The weight of SCI paper, SSCI paper, EI paper, International Conference paper (ICP), 1A paper, 1B paper, 2A paper, and 2B paper are set as 0.25, 0.25, 0.1, 0.04, 0.2, 0.1, 0.04, and 0.02 respectively according to the standard of scorecard for scientific research work in Zhejiang University of Finance and Economics. Finally, the SRA of an academic staff is defined as the sum of the number of published papers multiplied by corresponding weights.

3.3. Division of Attributes Interval

Because the values of attributes related to no. of published paper (e.g., SCI paper, and SSCI paper) are discrete, which may produce the rules with low support value. In order to obtain effective and useful rules, this study employs a feature selection package of R programming to obtain an effective division of these attributes.

Figure 1 is the division results of attributes related to no. of published paper. The X axis represents division of attributes interval, and the Y axis represents importance value. Green

boxes represent the division is acceptable, yellow box represent the division is pending, red boxes represent the division is rejective, and blue boxes represent a boundary and it can be ignored. The higher the value of importance is, the better the division is. Here, the symbol of SSCI>=2 represents the number of published SSCI paper is greater than or equal to 2. So do the other attributes. Because this operation is just for division of attributes interval, the division result of an attribute whose value of importance is the highest is reserved even if the result shows the rejection of this division.

In order to improve the accuracy and efficiency of association rule mining results, the data type is transformed to fit requirement of R programming. For example, the attribute "SCI" can be divided into two parts according to result of Figure 1, where the number of published SCI paper is lower than 4 can be marked as "0", the number of published SCI paper is greater than or equal to 4 can be marked as "1". The detailed divisions and representations of attributes are shown in Table 1. The partial preprocessing results of data are shown in Table 2.

3.4. Association Rule Mining

In this section, R programming and Apriori algorithm are employed to explore the rules between whether an academic staff can obtain NSFC within three years and personal attributes. In the process of association rule mining, three indexes including support, confidence, and lift are used to evaluate the rules.

Assume S is a set of items, X and Y satisfy the following conditions: $X \in S$, $Y \in S$, and $X \cap Y = \varnothing$. The symbol of $X=>Y$ represents an association rule. The support of rule $X=>Y$ is defined as the probability that X and Y occurred simultaneously. The confidence of rule $X=>Y$ is defined as the probability that Y occurred when X occurred. Lift is another index that is employed to find the rules between X and Y. When lift = 1, which represents that X and Y have no correlation. Therefore, this study only keeps the subset that the value of lift is higher than 1.

The mathematical representation of support, confidence, and lift of rule $X=>Y$ are shown in equations (1-3) respectively [17].

$$\text{Support}(X=>Y)=P(X \cap Y) \tag{1}$$

$$\text{Confidence}(X=>Y)=P(Y/X) \tag{2}$$

$$\text{Lift}(X=>Y)=\frac{P(Y/X)}{P(Y)} \tag{3}$$

Firstly, the itemset whose frequency are higher than or equal to predefined minimum support value is selected. Then the strong associated rules that satisfy the minimum support value and confidence value are generated based on the frequency itemset. In this study, the minimum value of support and confidence are set as 0.05 and 0.75 respectively.

4. Computational Experiments

Figure 2 is the scatter plot that represents the association rules of staff obtaining NSFC within three years. The X axis and Y axis represent the value of support and confidence respectively. Different points represent different association rules and different colors represent different values of lift. As shown in Figure 2, the support values of most association rules in this figure are lower than 0.3 and the confidence values of most association rules in this figure are higher than 0.75, but their lift values in this figure are higher than 1.2, which represents that most of staff have not obtained NSFC within three years after they are recruited to the university.

Figure 3 is the graph that represents the association rules of staff obtaining NSFC within three years. The size of circles is used to represent support value and the color of circles is used to represent lift value. The bigger the size of circle is, the higher the support value of rule is; the deeper the color of circle is, the higher the lift value of rule is. Because there are many rules of staff obtaining NSFC within three years, this study selects 10 rules of staff obtaining NSFC within three years and 10 rules of staff not obtaining NSFC within three years as examples.

Figure 3 shows that the rules of staff obtaining NSFC within three years have smaller support values than the rules of staff not obtaining NSFC within three years. However, the lift value for rules of staff obtaining NSFC within three years higher than the rules of staff not obtaining NSFC within three years. This phenomenon is probably produced due to the reason that most of staff are not obtained NSFC within three years. As shown in Figure 3, the staff with great SRA, especially, the staff who has published high quality paper (e.g., SCI paper) has higher probability to obtain NSFC within three years. Furthermore, the grade of PhD school has also an effect on the probability of academic staff to obtain NSFC within three years. The higher the grade of a staff's PhD school is, the easier for s/he to obtain NSFC within three years.

Table 3 is partial results of rules between obtained NSFC within three years and personal attributes. The first rule in Table 3 shows that the staff whose SRA greater than or equal to two and the number of published SCI paper is greater than or equal to 4 has a high probability to obtain NSFC within three years. The third rule in Table 3 shows that the staff whose PhD school belongs to C9 project has more chance to obtain NSFC within three years. This rule is probably caused by the reason that the higher the grade of a staff's PhD school is, the better research training the staff has been received.

Table 1. *The presentations of attributes.*

Attribute	Classification result
Grade of PhD school	0 = "foreign school", 1 = "C9" project, 2 = "985" project, 3 = "211" project, 4 = "other school"
Grade of undergraduate school	0 = "foreign school", 1 = "C9" project, 2 = "985" project, 3 = "211" project, 4 = "other school"
Grade of foreign PhD school	0 = "domestic school", 1 = "top 100 of Times or News", 2 = "top 150 of Times or News"

Attribute	Classification result
SRA	0 = "SRA < 2", 1 = "SRA ≥ 2"
SCI	0 = "the number of SCI papers < 4", 1 = "the number of SCI papers ≥ 4"
SSCI	0 = "the number of SSCI papers < 1", 1 = "the number of SSCI papers ≥ 1"
EI	0 = "the number of EI papers < 3", 1 = "the number of EI papers ≥ 3"
ICP	0 = "the number of ICP < 2", 1 = "the number of ICP ≥ 2"
1A	0 = "the number of 1A papers < 1", 1 = "the number of 1A papers ≥ 1"
1B	0 = "the number of 1B papers < 1", 1 = "the number of 1B papers ≥ 1"
2A	0 = "the number of 2A papers < 4", 1 = "the number of 2A papers ≥ 4"
National scientific research award	0 = "not obtained the National scientific research award", 1 = "obtained the National scientific research award"
Age	0 = "age ≤ 35", 1 = "35< age ≤ 45", 2 = "age ≥ 46"
Sex	0 = "female", 1 = "male"
Marital status	0 = "unmarried", 1 = "married", 2 = "unknown"
Obtained NSFC within three years	0 = "not obtained NSFC within three years", 1 = "obtained NSFC within three years"
Major category	0 = "Science", 1 = "Management", 2 = "Economics" 3 = "Law", 4 = "Engineering", 5 = "Literature", 6 = "Philosophy", 7 = "Pedagogy"

Table 2. *The partial rules of staff obtaining NSFC within three years.*

Obtain NSFC within three years	Sex	SRA	SCI	Grade of PhD school	National scientific research award
1	1	1	1	1	0
1	1	0	0	1	0
0	1	0	0	2	0
1	1	0	0	2	0
1	1	1	1	4	0
0	0	0	0	1	0
0	0	0	0	1	0
1	0	0	0	1	0
0	1	0	0	1	0
0	0	0	1	1	1
0	0	0	0	3	0
0	1	0	0	3	0
0	1	0	0	4	0
0	1	0	0	3	0
0	1	0	0	3	0

Table 3. *The partial rules of staff obtaining NSFC within three years.*

Rules	Support	Confidence	Lift
{SRA = 1, SCI = 1, ICP = 1} => {Obtained NSFC within three years = 1}	0.069767	1	2.388889
{SRA = 0, SCI = 0, SSCI = 0, 1A = 1, 1B = 0} => {Obtained NSFC within three years = 1}	0.052326	0.9	2.15
{Grade of PhD school = 1, 1A = 1, 2A = 0, Age = 0, Major category = 1, National scientific research award = 0} => {Obtained NSFC within three years = 1}	0.093023	0.8	1.911111
{Grade of undergraduate school = 4, SRA = 0, SCI = 0, Age = 1, Marital status = 1} => {Obtained NSFC within three years = 0}	0.168605	0.878788	1.511515
{SCI = 0, SSCI = 0, 1A = 0, 1B = 0, 2A = 0, Age = 1} => {Obtained NSFC within three years = 0}	0.127907	0.956522	1.645217
{SRA = 0, SSCI = 0, 1A = 0, 1B = 0, Age = 1, Sex = 1} => {Obtained NSFC within three years = 0}	0.145349	0.961538	1.653846

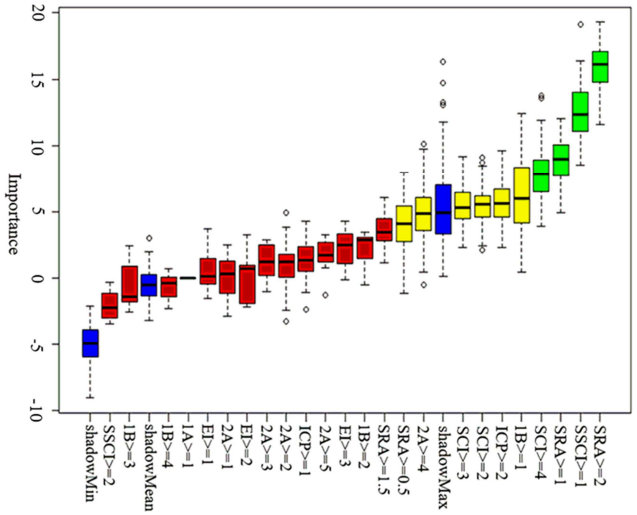

Figure 1. *The division of attributes related to no. of published paper.*

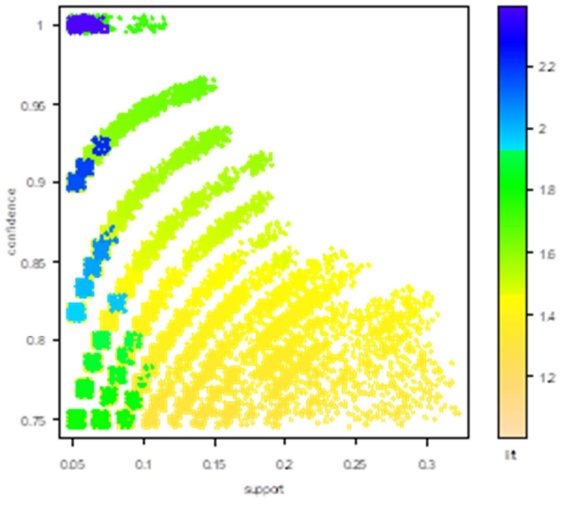

Figure 2. *Rules of staff obtaining NSFC within three years.*

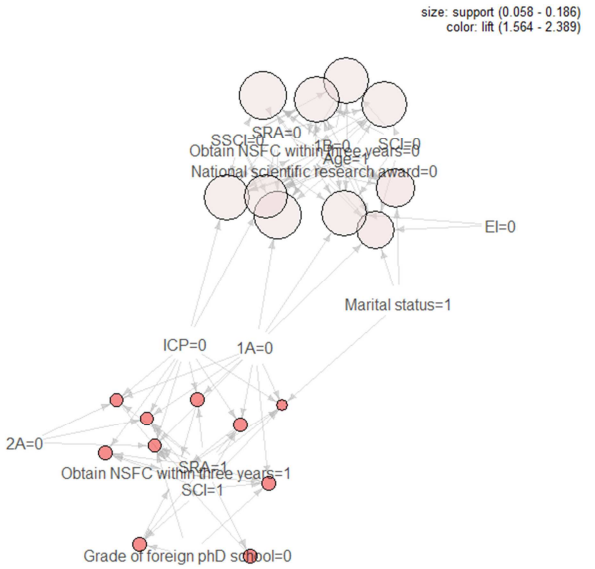

Figure 3. *Graph for rules of staff obtaining NSFC within three years.*

5. Conclusion

In this study, the method of association rule mining and Apriori algorithm are employed to explore the rules between whether an academic staff can obtain NSFC within three years and personal attributes. The results of association rule mining will help to colleges to introduce talents so that the academic levels of colleges can be improved. This study firstly collects the raw data of 245 academic staff in Zhejiang University of Finance and Economics. Then preprocesses the raw data and obtained the data that contains 172 academic staff with 17 personal attributes and the index indicating whether the NSFC can be obtained within three years after they are recruited to the university. The results of association rule mining have shown that publishing high quality paper such as SCI paper, and SSCI paper has an important effect on the probability of academic staff to obtain NSFC within three years. In other words, an academic staff with great SRA has more chance to obtain NSFC within three years. Besides, the grade of PhD school has also an effect on the probability of academic staff to obtain NSFC within three years. The higher the grade of a staff's PhD school is, the easier for the staff to obtain NSFC within three years. Therefore, the results of this study provide a reference for colleges to recruit talents introduction. The colleges should give priority to talents that have published high quality papers or graduated from high grade PhD school.

Currently, the problem of talent introduction becomes more and more complex. In this study, there are still some limitations that should be solved in further work. For example, because some academic staff haven't provided their personal information, this study only collects the data of 245 academic staff in Zhejiang University of Finance and Economics. The limited raw data has an effect on the accuracy of results. In addition, the association rule mining can combine with heuristic algorithm or other data mining technologies to obtain more accurate and efficient results in the future work.

References

[1] Chien C. F., and Chen L. F. (2008). "Data mining to improve personnel selection and enhance human capital: a case study in high-technology industry." Expert Systems with Applications, 34 (1): 280-290.

[2] Wu X. D., Kumar V., and Quinlan J. R. et al. (2008). "Top 10 algorithms in data mining." Knowledge and Information Systems, 14 (1): 1-37.

[3] Shi Y. (2010). "A dimension reduction approach using shrinking for multi-dimensional data analysis." International Journal of Intelligent Information Processing, 1 (2): 86-98.

[4] Lin W., Alvarez S. A., and Ruiz C. (2002). "Efficient adaptive-support association rule mining for recommender systems." Data Mining and Knowledge Discovery, 6 (1): 83-105.

[5] Qodmanan H. R., Nasiri M., and Minaei-Bidgoli B. (2011). "Multi objective association rule mining with genetic algorithm without specifying minimum support and minimum confidence." Expert Systems with Applications, 38 (1): 288-298.

[6] Huang C., Lu R., and Choo K. K. R. (2016). "Secure and flexible cloud-assisted association rule mining over horizontally partitioned databases." Journal of Computer and System Sciences, 89: 51-63.

[7] Beiranvand V., Mobasher-Kashani M., and Bakar A. A. (2014). "Multi-objective PSO algorithm for mining numerical association rules without a priori discretization." Expert Systems with Applications, 41 (9): 4259-4273.

[8] Gyenesei A. (2001). "A fuzzy approach for mining quantitative association rules." Acta Cybernetica, 15 (2): 305-320.

[9] Fung K. Y., Kwong C. K., and Siu K. W. M. et al. (2012). "A multi-objective genetic algorithm approach to rule mining for affective product design." Expert Systems with Applications, 39 (8): 7411-7419.

[10] Alatas B., Akin E., and Karci A. (2008). "MODENAR: multi-objective differential evolution algorithm for mining numeric association rules." Applied Soft Computing, 8 (1): 646-656.

[11] Agrawal R., Imieliński T., and Swami A. (1993). "Mining association rules between sets of items in large databases." In Proceedings of the ACM SIGMOD International Conference on Management of Data, May 26-28, Washington D. C., America, pp. 207-216.

[12] Guo Y., Wang M., and Li X. (2017). "Application of an improved apriori algorithm in a mobile e-commerce recommendation system." Industrial Management and Data Systems, 117 (2): 287-303.

[13] Li Q., Zhang Y. Y., and Kang H. Y. et al. (2017). "Mining association rules between stroke risk factors based on the apriori algorithm." Technology and Health Care, 25 (S1): 197-205.

[14] Guo Z., Chi D., and Wu J. et al. (2014). "A new wind speed forecasting strategy based on the chaotic time series modelling technique and the apriori algorithm." Energy Conversion and Management, 84: 140-151.

[15] Singh J., Ram H., and Sodhi D. J. (2013). "Improving efficiency of apriori algorithm using transaction reduction." International Journal of Scientific and Research Publications, 3 (1): 1-4.

[16] Yu H., Wen J., and Wang H. et al. (2011). "An improved apriori algorithm based on the Boolean matrix and Hadoop." Procedia Engineering, 15: 1827-1831.

[17] Agrawal R., and Srikant R. (1994). "Fast algorithms for mining association rules." In Proceedings of the 20th VLDB Conference, September 12-15, Santiago, Chile, pp. 487-499.